Convergence in human decision-making dynamics

Cao, Ming; Stewart, Andrew; Leonard, Naomi Ehrich

Published in:
Systems & Control Letters

DOI:
10.1016/j.sysconle.2009.12.002

Publication date:
2010

Systems & Control Letters 59 (2010) 87–97

Contents lists available at ScienceDirect

Systems & Control Letters

journal homepage: www.elsevier.com/locate/sysconle

Convergence in human decision-making dynamics

Ming Cao [a], Andrew Stewart [b], Naomi Ehrich Leonard [b,*]

[a] Faculty of Mathematics and Natural Sciences, University of Groningen, 9747 AG Groningen, The Netherlands
[b] Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08544, USA

ARTICLE INFO

Article history:
Received 26 June 2008
Received in revised form
25 November 2009
Accepted 7 December 2009
Available online 4 January 2010

Keywords:
Human decision making
Two-alternative forced-choice task
Win-stay, lose-switch model
Drift diffusion model
Robotic foraging
Explore vs. exploit

ABSTRACT

A class of binary decision-making tasks called the two-alternative forced-choice task has been used extensively in psychology and behavioral economics experiments to investigate human decision making. The human subject makes a choice between two options at regular time intervals and receives a reward after each choice; for a variety of reward structures, these experiments show convergence of the aggregate behavior to rewards that are often suboptimal. In this paper we present two models of human decision making: one is the Win-Stay, Lose-Switch (WSLS) model and the other is a deterministic limit of the popular Drift Diffusion (DD) model. With these models we prove the convergence of human behavior to the observed aggregate decision making for reward structures with matching points. The analysis is motivated by human-in-the-loop systems, where humans are often required to make repeated choices among finite alternatives in response to evolving system performance measures. We discuss application of the convergence result to the design of human-in-the-loop systems using a map from the human subject to a human supervisor.

1. Introduction

The superior ability of humans to handle the unexpected and to recognize patterns and extract structure from data makes human decision making invaluable to the performance of complex tasks in uncertain, changing environments. However, without computational aid, human decision makers may be overwhelmed by the many elements and the multitude of scales in complex, time-varying systems. This suggests an important role for automation, where fast, dedicated data processing and feedback responsiveness can be exploited. Indeed, there has been much research in the cooperative control of multi-agent robotic systems for automating cooperative tasks, see e.g., the collections [1–4]. However, fully automated decision making presents serious limitations to performance: automating optimal decisions in complex tasks in uncertain, time-varying settings will typically require solving intractable, nonlinear, stochastic, optimization problems. Even when good suboptimal decisions are designed into robotic decision makers, there will likely remain unanticipated, and thus poorly addressed, scenarios. Further, attempts to increase the versatility of the completely automated system can lead to unverifiable code.

The problem of integrating the efforts of humans and robots in demanding tasks has driven the rapidly growing fields of human–robot interaction and social robotics. For example, in [5] Simmons et al. have developed a framework and tools to coordinate robotic assembly teams with remote human operators for performing construction and assembly in hazardous environments. Their approach allows for adjustable autonomy where control of tasks can change between the human and the robotic team. To conserve human expertise, the robot team operates autonomously and only requests help from the human supervisor as needed. Using a similar notion of sliding autonomy, Kaupp and Makarenko [6] examine a human–robot communication system for information exchange in a navigation task. In [7] Steinfeld et al. define common metrics to evaluate human–robot interaction performance. Alami et al. [8] propose a robot control architecture that explicitly considers and manages its interaction with a human. The authors point out the critical challenge in representing the humans. Trafton and co-authors directly address this challenge by studying human–human interaction and then embedding the human behavior into the robots. For example, Trafton et al. [9] study videotape of astronauts-in-training performing cooperative assembly tasks and use the empirical data to inform cognitive models of perspective-taking. They embed these models into robots to improve their ability to work collaboratively with humans.

These works make clear that the profitable integration of human and robot decision-making dynamics should take advantage of the strengths of human decision makers as well as the strengths of robotic agents. They also make clear that a major challenge in

* Corresponding author. Tel.: +1 609 258 5129; fax: +1 609 258 6109.
E-mail addresses: M.Cao@rug.nl (M. Cao), arstewar@princeton.edu (A. Stewart), naomi@princeton.edu (N.E. Leonard).

0167-6911/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.sysconle.2009.12.002

achieving this goal is understanding how humans make decisions and what are their associated strengths and weaknesses. Since psychologists and behavioral scientists explore these very questions, a central tenet of our approach is to *leverage the experimental and modeling work of psychologists and behavioral scientists on human decision making*. To do this we seek commonality in the kinds of decisions humans make in complex tasks and the kinds of decisions humans make in psychology experiments. We focus here on a well-studied class of sequential binary decision making called the two-alternative forced-choice task [10–15]. In this task, the human subject in the psychology experiments chooses between two options at regular time intervals and receives a reward after each choice that depends on recent past decisions. Interestingly, these experiments show convergence of the aggregate behavior to rewards that are often suboptimal. This suboptimal behavior is of particular interest for what it can tell us about the dynamics of human-in-the-loop systems. Indeed, just like in the two-alternative forced-choice task, a human in the loop performing a supervisory task is often required to choose repeatedly between finite alternatives in response to an update on system performance that in turn depends on current and past decisions. Examples include human supervisory control of multiple unmanned aerial vehicles where the supervisor must decide between attending to targets and ensuring the safe return of vehicles [16] and human operator control of air traffic where the operator must decide between grounding or scheduling vehicles [17].

In this paper, we seek to formally describe the convergence behavior observed in experiments; the goal is to better understand the conditions of suboptimal human decision making and to provide a framework for designing improved human-in-the-loop systems. We present two models of human decision making in the two-alternative forced-choice task: the *Win-Stay, Lose-Switch* (WSLS) model and the *Drift Diffusion* (DD) model. With these models we prove convergence to matching behavior for reward structures with matching points. Intuitively, this implies that the decision maker converges to a neighborhood in the decision space of a decision sequence which results in a reward that stays roughly constant with each subsequent choice. These convergence results are consistent with experimental observations [10,18]. A partial version of our convergence results in the case of WSLS appears in [19]. We note that Montague and Berns [11] showed that the matching point is an attracting point in the DD model in the case that a couple of simplifying assumptions hold. In [20] Vu and Morgansen perform an analysis of the WSLS model in a related context using finite state machines.

As one possible application of the convergence results, we introduce a human-supervised collective robotic foraging problem, where a group of robots in the field moves around and collects a distributed resource and a human supervisor repeatedly makes a decision to assign the role of each of the robots, either to be an explorer or an exploiter of resource. The human and robots work as a team to maximize resource collected: the robots follow efficient, collaborative exploring and exploiting strategies and the human tries to make the best allocation decisions.

We review studies of the two-alternative forced-choice task in Section 2. In Section 3, we present the WSLS and DD decision-making models. We prove convergence to matching for each of these models in Section 4. In Section 5 we present, and investigate with a simulation study, a map from the decision-making task of a human supervisor of a robotic foraging team to the two-alternative forced-choice task. We make final remarks in Section 6.

2. Two-alternative force-choice tasks

Real-world decision-making problems are difficult to study since the reward for a decision usually depends in a nontrivial way on the decision history. Many studies have considered decision-reward relationships that are fixed; however, these have a limited value in addressing problems associated with complex, time-varying tasks like the ones motivating this paper. Here, we briefly review a class of decision-making tasks called the *two-alternative forced-choice task*, where reward depends on past decisions.

Montague et al. [14,11] introduced a dynamic economic game with a series of decision-reward relationships that depend on a subject's decision history. The human subject is faced with a two-alternative sequential choice task. Choices of either A or B are made sequentially (forced within regular time intervals) and a reward for each decision is administered directly following the choice. Without knowing the reward structure, the human subject tries to maximize the total reward (sum of individual rewards).

Two reward structures frequently considered are the *matching shoulders* (Fig. 1(a)) and *rising optimum* (Fig. 1(b)). As shown in these figures, the reward f_A for choosing A and the reward f_B for choosing B are determined as a function of the proportion y of times A was chosen in the previous N trials ($y = \#A\text{'s}/N$). $N = 20$ and $N = 40$ are typical choices in experiments. In Fig. 1(a), if the subject always chooses A, the reward drops to 0.2. Subsequently, if B is chosen, the reward jumps up close to 0.9. However, continued choices of B lead to declining reward. The average reward, plotted as a dashed line on each figure, is computed as $yf_A(y)+(1-y)f_B(y)$. The optimal strategy is the one that maximizes the average reward curve.

Experimental studies show how humans perform at this task with specific reward structures. A great number of experiments consider the matching shoulders and rising optimum tasks and these experiments have shown that human subjects tend, on average, to adopt choice sequences y that bring them close to the *matching point* of the reward curves (intersection of reward curves) [10,18,14,11]. Herrnstein [10,18] explains that this is reasonable since near the matching point the reward for choosing A or B is about the same. However, this implies that humans do not necessarily converge on the optimal strategy, since the matching point does not necessarily correspond to the optimal average reward. In Fig. 1(a), the optimal reward is to the right of the matching point. In Fig. 1(b), the optimal strategy corresponds to choosing option A every time. The reward at the matching point is significantly suboptimal and to reach the optimum the subject may first have to endure very low rewards.

To model the two-alternative forced-choice task we define the following notation. Let $x_1(t) \in \{A, B\}$ denote the decision for the binary choice A or B at time t and let $x_i(t) = x_1(t - i + 1)$, $i = 2, \ldots, N$, denote the $N - 1$ most recent decisions of the finite past. Equivalently, we have

$$x_i(t + 1) = x_{i-1}(t), \quad i = 2, \ldots, N, \ t = 0, 1, 2, \ldots. \tag{1}$$

Let y denote the proportion of choice A in the last N decisions, i.e.,

$$y(t) = \frac{1}{N}\sum_{i=1}^{N} \delta_{iA}(t) \tag{2}$$

where

$$\delta_{iA}(t) = \begin{cases} 1 & \text{if } x_i(t) = A \\ 0 & \text{if } x_i(t) = B. \end{cases}$$

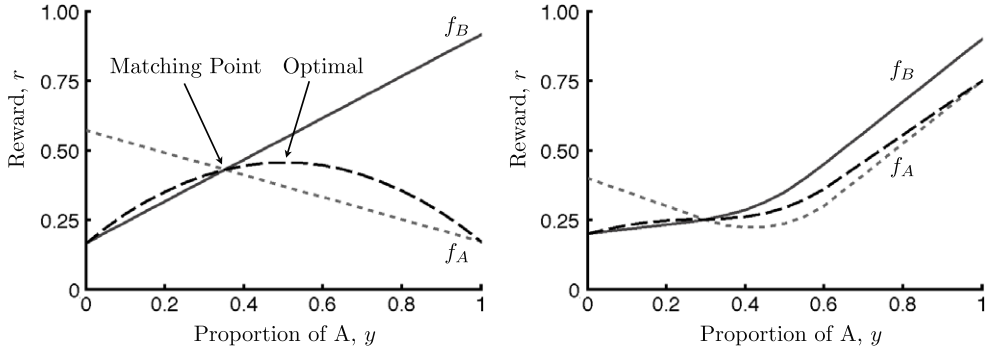
Note that y can only take values from a finite set $\mathcal{Y}$ of $N+1$ discrete values:

$$\mathcal{Y} = \{j/N, \ j = 0, 1, \ldots, N\}.$$

The reward at time t is given by

$$r(t) = \begin{cases} f_A(y(t)) & \text{if } x_1(t) = A \\ f_B(y(t)) & \text{if } x_1(t) = B. \end{cases} \tag{3}$$

Thus, the two-alternative forced-choice task can be modeled as an N-dimensional, discrete-time dynamical system, described completely by Eqs. (1)–(3) where $x_i(t)$, $1 \le i \le N$, is the state of the system and $y(t)$ is the output of the system.



(a) Matching shoulders. (b) Rising optimum.

Fig. 1. Reward curves. The dotted line depicts f_A, the reward for choice A. The solid line depicts f_B, the reward for choice B. The dashed line is the average value of the reward. Each is plotted against the proportion of choice A made in the last N decisions.

3. Decision-making models

The matching tendency in humans and animals was first identified by Herrnstein [10,18], whose related work has been influential in the quantitative analysis of behavior and mathematical psychology. However, few mathematically provable results on matching behavior have been obtained and reported. This is, in part, due to the difficulty in modeling the dynamics of human and animal decision making. Several models have been proposed to describe the dynamics of human decision making, and in this paper we analyze two of them. One is the *Win-Stay, Lose-Switch* (WSLS) model, also known as *Win-Stay, Lose-Shift*, which has been used in psychology, game theory, statistics and machine learning [21,22]. The other is a deterministic limit of a popular stochastic reinforcement learning model, called the *Drift Diffusion* (DD) model, which was introduced by Egelman et al. [14] and further studied in [11,12]. It should be noted that many of the other models which have been proposed are, in fact, equivalent to the DD model [23].

3.1. WSLS model

The subject's decision dynamics may be affected by all the decisions and rewards in the last N trials. In fact, a goal of the studies of two-alternative forced-choice tasks is to determine the decision-making mechanism through experiment and use behavioral and neurobiological arguments to justify the likelihood of the mechanism. The WSLS model assumes that decisions are made with information from the rewards of the previous two choices only and that a switch in choice is made when a decrease in reward is experienced. That is, the subject repeats the choice from time t at time $t+1$ if the reward at time t is greater than or equal to that at time $t-1$; otherwise, the opposite choice is used at time $t+1$:

$$x_1(t+1) = \begin{cases} x_1(t) & \text{if } r(t) \geq r(t-1); \\ \bar{x}_1(t) & \text{otherwise,} \end{cases} \qquad t = 1, 2, 3, \ldots \tag{4}$$

where $\bar{\cdot}$ denotes the "not" operator; i.e., if $x_1(t) = A$ ($x_1(t) = B$), then $\bar{x}_1(t) = B$ ($\bar{x}_1(t) = A$).

3.2. A deterministic limit of the DD model

In the context of multi-choice decision-making dynamics, a one-dimensional drift diffusion process can be described by a stochastic differential equation [24,15,25]:

$$\mathrm{d}z = \alpha \mathrm{d}t + \sigma \mathrm{d}W, \qquad z(0) = 0 \tag{5}$$

where z represents the accumulated evidence in favor of one choice of interest, α is the drift rate representing the signal intensity of the stimulus acting on z and $\sigma\,\mathrm{d}W$ is a Wiener process with standard deviation σ which is the diffusion rate representing the effect of white noise. Now consider the two-alternative forced-choice task with choices A and B. The drift rate α, as described in [14,12], is determined by a subject's anticipated rewards for a decision of A or B, denoted ω_A and ω_B.

Take z to be the accumulated evidence for choice A less the accumulated evidence for choice B. Then on each trial a choice is made when $z(t)$ *first* crosses the predetermined thresholds $\pm\nu$. In which case, if ν is crossed choice A is made and if $-\nu$ is crossed choice B is made. For such drift diffusion processes, as pointed out in [15], it can be computed using tools developed in [23] that the probability of choosing A is

$$p_A(t) = \frac{1}{1 + \mathrm{e}^{-\mu(\omega_A(t) - \omega_B(t))}} \tag{6}$$

where μ is determined by the threshold-to-drift ratio $\frac{\nu}{\alpha}$, and $\omega_A - \omega_B$ is determined by the signal-to-noise ratio $\frac{\alpha}{\sigma}$.

To gain insight into the mechanics of the DD model, we choose a specific set of relevant parameters. Specifically, we study the deterministic limit of the decision rule (6) deduced from the DD model by letting μ in (6) go to infinity. Then at time $t > 0$, the subject chooses A if $\omega_A(t) > \omega_B(t)$ and B if $\omega_A(t) < \omega_B(t)$. In the event that $\omega_A = \omega_B$, we assume that humans are of an explorative nature, and thus the subject uses the opposite of the last choice. To summarize,

$$x_1(t) = \begin{cases} A & \text{if } \omega_A(t) > \omega_B(t) \\ B & \text{if } \omega_A(t) < \omega_B(t) \\ \bar{x}_1(t-1) & \text{if } \omega_A(t) = \omega_B(t) \end{cases} \qquad t = 1, 2, 3, \ldots \tag{7}$$

where as in (4), $\bar{\cdot}$ denotes the "not" operator.

We are now left with modelling the transition of the anticipated rewards ω_A and ω_B. Using data collected in neurobiological studies of the role of dopamine neurons in coding for reward prediction error [26], and guided by temporal difference learning theory [27], the following difference equations have been proposed to describe the update of ω_A and ω_B. Let $Z(t) \in \{A, B\}$ be the choice made at time t, then

$$\omega_{Z(t)}(t+1) = (1-\lambda)\omega_{Z(t)}(t) + \lambda r(t) \tag{8}$$

$$\omega_{\bar{Z}(t)}(t+1) = \omega_{\bar{Z}(t)}(t) \quad t = 0, 1, 2, \ldots \tag{9}$$

where $r(t)$ is the reward at time t. Here, $\lambda \in [0, 1]$ is called the *learning rate*, which reflects how the anticipated reward of choice $Z(t)$ at $t+1$ is affected by its value at t.

In the update model (8) and (9), referred to as the *standard model* in the sequel, when a choice of Z is made the value of $\omega_{\bar{Z}}$ remains unchanged because without memory no reward information for $\bar{Z}$ is available. A more sophisticated update model, called the *eligibility trace model*, is constructed in [12]. It takes

into account the effect of memory by updating both ω_A and ω_B continually. The *eligibility trace* can be interpreted as a description of how psychological perception of information refreshes or decays in response to whether or not an external stimulus is enforced. The eligibility traces (as presented in [12]) denoted by $\phi_A(t)$ and $\phi_B(t)$ for choices A and B respectively, evolve according to

$$\phi_{Z(t)}(t+1) = 1 + \phi_{Z(t)}(t)\mathrm{e}^{-\frac{1}{\tau}} \tag{10}$$

$$\phi_{\bar{Z}(t)}(t+1) = \phi_{\bar{Z}(t)}(t)\mathrm{e}^{-\frac{1}{\tau}} \tag{11}$$

with initial values $\phi_A(0) = \phi_B(0)$, where $\tau > 0$ is a parameter that determines the decaying effects of memories.

With the eligibility traces included, ω_A and ω_B are updated according to

$$\omega_A(t+1) = \omega_A(t) + \lambda[r(t) - \omega_{Z(t)}(t)]\phi_A(t) \tag{12}$$

$$\omega_B(t+1) = \omega_B(t) + \lambda[r(t) - \omega_{Z(t)}(t)]\phi_B(t) \tag{13}$$

where the eligibility traces ϕ_A and ϕ_B act as time-varying weighting factors. When τ is chosen to be small, the update rule in the eligibility trace model (12) and (13) reduces to that in the standard model (8) and (9).

To analyze the impact of the dynamics of the eligibility traces on the evolution of ω_A and ω_B, we discretize the eligibility traces ϕ_A and ϕ_B. We set the learning rate λ in (12) and (13) to be its maximum value ($\lambda = 1$) which corresponds to the current reward having the strongest possible influence on the subject. Then

$$\omega_A(t+1) = \omega_A(t) + [r(t) - \omega_{Z(t)}(t)]\phi_A(t) \tag{14}$$

$$\omega_B(t+1) = \omega_B(t) + [r(t) - \omega_{Z(t)}(t)]\phi_B(t). \tag{15}$$

We discretize ϕ_A and ϕ_B as follows. For $\sigma \in \{A, B\}$, if σ is the last choice made, we set the value of ϕ_σ to be saturated at one because the impact of the current reward has been accounted for by setting λ to be its maximum value. We let ϕ_σ decay to zero once the opposite choice $\bar{\sigma}$ has been chosen consecutively. This corresponds to the situation when, without an external stimulus, the memory fades quickly. When a switch in choice is made (from σ at time $t-1$ to $\bar{\sigma}$ at time t), let the fresh memory of the unchosen alternative, ϕ_σ, be a small, positive number $\epsilon \in (0, 1)$. Then ϕ_A and ϕ_B take values in $\{0, \epsilon, 1\}$ and evolve according to

$$\phi_\sigma(t) = \begin{cases} 1 & \text{if } \sigma = x_1(t) \\ \epsilon & \text{if } \sigma = \bar{x}_1(t) = x_1(t-1) \\ 0 & \text{if } \sigma = \bar{x}_1(t) = \bar{x}_1(t-1) \end{cases} \quad t = 1, 2, 3, \ldots. \tag{16}$$

The resulting model is a deterministic limit of the DD with eligibility traces.

4. Convergence analysis

In this section, we give a rigorous analysis of the dynamics of human performance in games with matching shoulders rewards. It is shown that for both of the human decision-making models, the proportion y of choice A converges to a neighborhood of the matching point. It should be noted that convergence also applies for reward structures that contain the matching shoulders structure locally (as in the rising optimum example of Fig. 1(b)).

Denote by y^* the value of y at the matching point, i.e., the intersection of the two curves f_A and f_B. We consider the generic case when

$$y^* \notin \mathcal{Y}, \tag{17}$$

i.e., y^* is not an integer multiple of $1/N$. In the non-generic case, when $y^* \in \mathcal{Y}$, a tighter convergence result applies. Let y^l denote the greatest element in $\mathcal{Y}$ that is smaller than y^* and let y^u denote the smallest element in $\mathcal{Y}$ that is greater than y^*. Let $y^{l'} = y^l - 1/N$ and $y^{u'} = y^u + 1/N$. Define

$$\mathcal{L} \triangleq [y^l, y^u] \quad \text{and} \quad \mathcal{L}' \triangleq [y^{l'}, y^{u'}].$$

So that $\mathcal{L}'$ is well defined, let $1/N < y^* < (N-1)/N$ and $N \geq 3$.

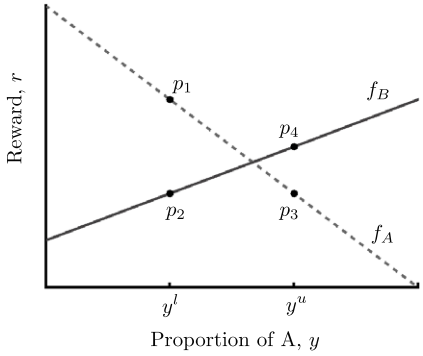


Fig. 2. Points p_1, p_2, p_3, and p_4 used to examine trajectories around the matching point.

4.1. Convergence of the WSLS model

In this section, we analyze the convergence behavior of the system (1)–(4). We consider reward curves such that f_A decreases monotonically and f_B increases monotonically with an increasing y, i.e.,

$$\frac{\mathrm{d}}{\mathrm{d}y} f_A(y) < 0, \qquad \frac{\mathrm{d}}{\mathrm{d}y} f_B(y) > 0, \quad \forall y \in [0, 1]. \tag{18}$$

This includes the linear matching shoulders curves of Fig. 1(a) as well as a more general class of nonlinear reward curves. It also includes the rising optimum reward curves of Fig. 1(b) locally about the matching point. We prove both a local and a global convergence result to the matching point. The local result is applicable to reward structures that have a local matching point and satisfy (18) for y in a neighborhood of y^*. This includes the rising optimum reward curves of Fig. 1(b). To formally preclude limit cycles about points other than the matching point it is necessary to require that

$$\frac{1}{3} \leq y^* \leq \frac{2}{3}. \tag{19}$$

The convergence results apply for general $N \geq 6$. For lower values of N the system degenerates and the output y may converge to 0 or 1. The linear curves used in the experiments [11] satisfy the conditions (17), (18) and (19), so the analysis in this section provides an analytical understanding of human decision-making dynamics in two-alternative forced-choice tasks of the same type.

4.1.1. Local convergence

The following result describes the oscillating behavior of $y(t)$ near y^*.

Theorem 1. *For system (1)–(4) satisfying conditions (17)–(19), if $y(t_1) \in \mathcal{L}$ for some $t_1 > 0$, then $y(t) \in \mathcal{L}'$ for all $t \geq t_1$.*

Before we prove Theorem 1, let us first examine a typical trajectory starting at time $t = t_1$ with $y(t_1) \in \mathcal{L}$. Consider the matching shoulders reward structure shown in Fig. 2 as an example. We bring attention to four points on the reward curves. As shown in Fig. 2 (which represents one possible configuration of these four points for general matching shoulders reward curves), we denote $p_1 = (y^l, f_A(y^l))$, $p_2 = (y^l, f_B(y^l))$, $p_3 = (y^u, f_A(y^u))$ and $p_4 = (y^u, f_B(y^u))$.

Suppose we are given a set of initial conditions $y(t_1) = y^u$, $x_1(t_1) = A$, $x_N(t_1) = B$ and suppose $x_1(t_1+1) = B$. Then $y(t_1+1) = y(t_1) = y^u$ and the reward $r(t_1+1) = f_B(y^u) > f_A(y^u) = r(t_1)$. In view of (4), we know that $x_1(t_1+2) = B$. If $x_N(t_1+1) = A$, then $y(t_1+2) = y(t_1+1) - 1/N = y^l$ and $r(t_1+2) = f_B(y^l) < f_B(y^u) = r(t_1+1)$. Again by (4), it must be true that $x_1(t_1+3) = A$. Suppose $x_N(t+2) = B$, then $y(t_1+3) = y^u$.

To track the above system trajectory in Fig. 2 for $t_1 \le t \le t_1+3$, one can find that the trajectory moves from p_3, to p_4, to p_2 and back to p_3. Hence, one may conjecture that once $y(t)$ enters $\mathcal{L}$, it will stay in $\mathcal{L}$. However, this is not the case. Consider a counterexample, namely in Fig. 2 the system trajectory again starts at p_3. However, let $x_N(t_1) = B$ and $x_1(t_1+1) = A$. Then $y(t_1+2) = y^u + 1/N \notin \mathcal{L}$. Although $\mathcal{L}$ is not an invariant set for $y(t)$, trajectories of $y(t)$ starting in $\mathcal{L}$ will always remain in $\mathcal{L}'$. This is an intuitive interpretation of Theorem 1.

To prove Theorem 1, we first prove the following four lemmas.

Lemma 1. *For system (1)–(4), with conditions (17)–(19) satisfied, if $x_1(t_1) = A, x_1(t_1+1) = A$ and $y(t_1) < 1$ for some $t_1 \ge 0$, then there exists $0 \le \tau \le N$ such that $y(t) = y(t_1)$ for $t_1 \le t \le t_1+\tau$ and $y(t_1+\tau+1) = y(t_1) + 1/N$.*

Proof of Lemma 1. If $x_N(t_1) = B$, then $y(t_1+1) = y(t_1) + 1/N$. So the conclusion holds for $\tau = 0$. On the other hand, if $x_N(t_1) = A$, then $y(t_1+1) = y(t_1)$ and $r(t_1+1) = f_A(y(t_1+1)) = f_A(y(t_1)) = r(t_1)$. According to (4), $x_1(t_1+2) = A$. In fact the choice of A will be repeatedly chosen as long as the value of x_N remains A. However, since $y(t_1) < 1$, there must exist $0 \le \tau < N$ such that $x_N(t) = A$ for $t_1 \le t \le t_1+\tau$ and $x_N(t_1+\tau+1) = B$. Accordingly, the conclusion holds. □

One can prove the following lemma, the counterpart to Lemma 1, with a similar argument.

Lemma 2. *For system (1)–(4), with conditions (17)–(19) satisfied, if $x_1(t_1) = B, x_1(t_1+1) = B$ and $y(t_1) > 0$ for some $t_1 \ge 0$, then there exists $0 \le \tau \le N$ such that $y(t) = y(t_1)$ for $t_1 \le t \le t_1+\tau$ and $y(t_1+\tau+1) = y(t_1) - 1/N$.*

Now we further study behavior of the system when its trajectory is on the left of the matching point y^*.

Lemma 3. *For system (1)–(4), with conditions (17)–(19) satisfied, if $y(t_1) < y^*$ and $y(t_1+1) = y(t_1) - 1/N > 0$ for some $t_1 \ge 0$, then there exists $0 \le \tau \le N$ such that*

$$y(t) = y(t_1) - 1/N \quad \text{for } t_1 \le t \le t_1+\tau \tag{20}$$

and

$$y(t_1+\tau+1) = y(t_1). \tag{21}$$

Proof of Lemma 3. We find it convenient to prove this lemma by labeling the following four points: $s_1 = (y(t_1), f_A(y(t_1)))$, $s_2 = (y(t_1), f_B(y(t_1)))$, $s_3 = (y(t_1)-1/N, f_A(y(t_1)-1/N))$, $s_4 = (y(t_1)-1/N, f_B(y(t_1)-1/N))$, as shown in Fig. 3.

We denote the reward values at these four points by $r|_{s_i}$, $i = 1, \ldots, 4$. Then $r(t_1) = r|_{s_1}$ or $r|_{s_2}$. Since $y(t_1+1) < y(t_1)$, it must be true that $x_1(t_1+1) = B$, then $r(t_1+1) = r|_{s_4}$. Since $r|_{s_4} < r|_{s_2} < r|_{s_1}$, we know $x_1(t_1+2) = A$. So at t_1+2, the system trajectory moves from s_4 to either s_1 or s_3. If the former is true, the conclusion holds for $\tau = 2$. If the latter is true, since $r|_{s_3} > r|_{s_4}$, it follows that $x_1(t_1+3) = A$. By applying Lemma 1, we know (20) and (21) hold. □

Similarly, we consider the situation when $y(t_1) > y^*$ and $y(t_1+1) = y(t_1) + 1/N < 1$ for some $t_1 \ge 0$. Denote four points: $r_1 = (y(t_1), f_A(y(t_1)))$, $r_2 = (y(t_1), f_B(y(t_1)))$, $r_3 = (y(t_1)+1/N, f_A(y(t_1)+1/N))$ and $r_4 = (y(t_1)+1/N, f_B(y(t_1)+1/N))$. Using the fact that $r|_{r_3} < r|_{r_1} < r|_{r_2}$ and a similar argument as that in the proof of Lemma 3, we can prove the following result.

Lemma 4. *For system (1)–(4), with conditions (17)–(19) satisfied, if $y(t_1) > y^*$ and $y(t_1+1) = y(t_1) + 1/N < 1$ for some $t_1 \ge 0$, then there exists $0 \le \tau \le N$ such that*

$$y(t) = y(t_1) + 1/N \quad \text{for } t_1 \le t \le t_1+\tau \tag{22}$$

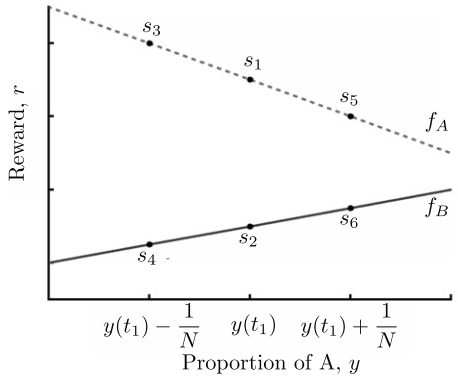


Fig. 3. Points s_1, s_2, s_3, s_4, s_5, and s_6 used in the proofs of Lemmas 3 and 7.

and

$$y(t_1+\tau+1) = y(t_1). \tag{23}$$

Now we are in a position to prove Theorem 1.

Proof of Theorem 1. If $y(t) \in \mathcal{L}$ for all $t \ge t_1$, then the conclusion holds trivially. Now suppose this is not true. Let $t_2 > t_1$ be the first time for which $y(t) \notin \mathcal{L}$. Then it suffices to prove the claim that the trajectory of $y(t)$ starting at $y(t_2)$ stays at $y(t_2)$ for a finite time and then enters $\mathcal{L}$. Note that $y(t_2)$ equals either $y^l - 1/N$ or $y^u + 1/N$. Suppose $y(t_2) = y^l - 1/N$, then the claim follows directly from Lemma 3; if on the other hand, $y(t_2) = y^u + 1/N$, then the claim follows directly from Lemma 4. □

4.1.2. Global convergence

Theorem 1 gives the convergence analysis in the neighborhood $\mathcal{L}$ of the matching point y^*. Our next step is to present the global convergence analysis for the system (1)–(4). It is easy to check that if the system starts with the initial condition $y(0) = 0$ and $x_1(1) = B$ or the initial condition $y(0) = 1$ and $x_1(1) = A$, then the trajectory of $y(t)$ will stay at its initial location. It will also be shown that when $y^* < \frac{1}{3}$ or $y^* > \frac{2}{3}$, a limit cycle of period three not containing y^* may appear. Thus it is necessary that condition (19) is satisfied, i.e., $\frac{1}{3} \le y^* \le \frac{2}{3}$. In what follows we show that if the trajectory of $y(t)$ starts in $(0, 1)$ and conditions (17), (18) and (19) are satisfied, then the trajectory always enters $\mathcal{L}$ after a finite time.

Proposition 1. *For any initial condition of the system (1)–(4) satisfying $0 < y(0) < 1$ with conditions (17)–(19) satisfied, there is a finite time $T > 0$ such that $y(T) \in \mathcal{L}$.*

To prove Proposition 1, we need to prove the following four lemmas.

Lemma 5. *For system (1)–(4), with conditions (17)–(19) satisfied, if $y(t_1) < y^*$, $y(t_1+1) = y(t_1)$ and $x_1(t_1+1) \ne x_1(t_1)$ for some $t_1 \ge 0$, then there exists a finite $\tau > 0$ such that*

$$y(t_1+\tau) = y(t_1) + 1/N. \tag{24}$$

Proof of Lemma 5. There are two cases to consider. (a) Suppose $x_1(t_1+1) = A$ and $x_1(t_1) = B$. Since $y(t_1+1) = y(t_1) < y^*$, we know $r(t_1+1) = f_A(y(t_1+1)) = f_A(y(t_1)) > f_B(y(t_1)) = r(t_1)$, so $x_1(t_1+2) = A$. Then the conclusion follows from Lemma 1. (b) Now suppose instead $x_1(t_1+1) = B$ and $x_1(t_1) = A$. Again since $y(t_1+1) = y(t_1) < y^*$, we know $r(t_1+1) = f_B(y(t_1+1)) = f_B(y(t_1)) < f_A(y(t_1)) = r(t_1)$, so $x_1(t_1+2) = A$. As a result, either $y(t_1+2) = y(t_1) + 1/N$ or $y(t_1+2) = y(t_1+1)$. If the former is true, then the conclusion holds for $\tau = 2$; if the latter is true, then the discussion reduces to that in (a). □

Using a similar argument, one can prove the following lemma, which is the counterpart to Lemma 5.

Lemma 6. *For system (1)–(4), with conditions (17)–(19) satisfied, if* $y(t_1) > y^*$, $y(t_1+1) = y(t_1)$ *and* $x_1(t_1+1) \neq x_1(t_1)$ *for some* $t_1 \geq 0$, *then there exists a finite* $\tau > 0$ *such that*

$$y(t_1+\tau) = y(t_1) - 1/N. \tag{25}$$

Now we show that the system will approach the matching point.

Lemma 7. *For system (1)–(4), with conditions (17)–(19) satisfied, if* $0 < y(t_1) < y^l$ *and* $y(t_1+1) = y(t_1) - 1/N$ *for some* $t_1 \geq 0$, *then there exists a finite* $\tau > 0$ *such that*

$$y(t_1+\tau) = y(t_1) + 1/N. \tag{26}$$

Proof of Lemma 7. Denote the six points $s_1 = (y(t_1), f_A(y(t_1)))$, $s_2 = (y(t_1), f_B(y(t_1)))$, $s_3 = (y(t_1) - 1/N, f_A(y(t_1) - 1/N))$, $s_4 = (y(t_1) - 1/N, f_B(y(t_1) - 1/N))$, $s_5 = (y(t_1) + 1/N, f_A(y(t_1) + 1/N))$ and $s_6 = (y(t_1) + 1/N, f_B(y(t_1) + 1/N))$, as shown in Fig. 3. Since $y(t_1+1) < y(t_1)$, it must be true that $x_1(t_1+1) = B$. If $x_1(t_1) = A$, we know from t_1 to t_1+1, the system trajectory moves from s_1 to s_4. Since $r(t_1+1) = r|_{s_4} < r|_{s_1} = r(t_1)$, we know $x_1(t_1+2) = A$. Then at t_1+2, the trajectory moves to either s_1 or s_3. We discuss these two cases separately.

(a) If at t_1+2 the trajectory moves to s_1, since $r|_{s_1} > r|_{s_4}$, it follows that $x_1(t_1+3) = A$. In view of Lemma 1, the conclusion holds.

(b) If at t_1+2, the trajectory moves to s_3, since $r|_{s_3} > r|_{s_4}$, we know $x_1(t_1+3) = A$. From Lemma 1, there exists a finite time $t_2 < N$ at which the trajectory moves from s_3 to s_1. Because $r|_{s_1} < r|_{s_3}$, we have $x_1(t_2+1) = B$. Then at time t_2+1, the trajectory moves to either s_2 or s_4. Now we discuss two sub-cases. (b_1) Suppose the former is true, that the trajectory goes to s_2. The conclusion follows directly from Lemma 5. (b_2) Suppose the latter is true, that the trajectory goes to s_4. Because $r|_{s_4} < r|_{s_1}$, $x_1(t_2+2) = A$. Then $y(t)$ will remain strictly less than $y(t_1) + 1/N$ if a cycle of $s_4 \to s_3 \to s_1 \to s_4$ is formed. In fact, from the analysis above, this is the only potential scenario in case (b_2) where $y(t) < y(t_1) + 1/N$ for all $t \geq t_1$. Were such a cycle to appear, A would be chosen at least twice as often as B. However, because $y(t_1) < y^l = y^* - 1/N < \frac{1}{3}$, it must be true that the proportion of A in $x_i(t_1)$, $1 \leq i \leq N$, is less than $\frac{1}{3}$. Thus such a cycle can never happen. So the conclusion also holds for the sub-case (b_2).

If on the other hand, $x_1(t_1) = B$, we know from t_1 to t_1+1, the system trajectory moves from s_2 to s_4. Since $r|_{s_4} < r|_{s_2}$, we know $x_1(t_1+2) = A$. So at t_1+2, the trajectory moves to s_3 or s_1. If the former is true, the discussion reduces to ruling out the possibility of forming a cycle of $s_4 \to s_3 \to s_1 \to s_4$ which we have done in (b_2). Otherwise, if the latter is true, since $r|_{s_1} > r|_{s_4}$, we know $x_1(t_1+3) = A$. From Lemma 1 we know there exists a finite time t_3 at which $y(t_3) = y(t_1) + 1/N$, and thus the conclusion holds for $\tau = t_3 - t_1$.

Combining the above discussions, we conclude that the proof of Lemma 7 is complete. □

Using Lemmas 2, 6 and a similar argument as in the proof of Lemma 7, one can prove the following lemma which is the counterpart to Lemma 7.

Lemma 8. *For system (1)–(4), with conditions (17)–(19) satisfied, if* $y^u < y(t_1) < 1$ *and* $y(t_1+1) = y(t_1) + 1/N$ *for some* $t_1 \geq 0$, *then there exists a finite* $\tau > 0$ *such that*

$$y(t_1+\tau) = y(t_1) - 1/N. \tag{27}$$

Now we are in a position to prove Proposition 1.

Proof of Proposition 1. For any $0 < y(0) < 1$, either $y(1) = y(0) + 1/N$, or $y(1) = y(0)$, or $y(1) = y(0) - 1/N$. We will discuss these three possibilities in each of two cases. First consider the case where $y(0) < y^l$. If $y(1) = y(0) - 1/N$, according to Lemma 7, there is a finite time t_1 for which $y(t_1) > y(0)$. If $y(1) = y(0)$ and $x_1(1) \neq x_1(0)$, according to Lemma 5, there is a finite time t_2 for which $y(t_2) > y(0)$. If $y(1) = y(0)$ and $x(1) = x(0) = A$, according to Lemma 1, there is a finite time t_3 for which $y(t_3) > y(0)$. If $y(1) = y(0)$ and $x(1) = x(0) = B$, according to Lemma 2, there is a finite time $\bar{t}_4$ for which $y(\bar{t}_4 - 1) = y(0)$ and $y(\bar{t}_4) = y(\bar{t}_4 - 1) - 1/N$. Then according to Lemma 7, there is a finite time t_4 for which $y(t_4) > y(0)$. So for all possibilities of $y(1)$ there is always a finite time $\bar{t} \in \{1, t_1, t_2, t_3, t_4\}$ for which $y(\bar{t}) > y(0)$. Using this argument repeatedly, we know that there exists a finite time T_1 at which $y(T_1) = y^l \in \mathcal{L}$. Now consider the other case where $y(0) > y^u$, then using similar arguments, one can check that there exists a finite time T_2 for which $y(T_2) = y^u \in \mathcal{L}$. Hence, we have proven the existence of T which lies in the set $\{T_1, T_2\}$. □

Combining the conclusions in Theorem 1 and Proposition 1, we have proven the following theorem, which describes the global convergence property of $y(t)$.

Theorem 2. *For any initial condition of the system (1)–(4) satisfying* $0 < y(0) < 1$ *with conditions (17)–(19) satisfied, there exists a finite time* $T > 0$ *such that for any* $t \geq T$, $y(t) \in \mathcal{L}'$.

4.2. Convergence of the DD model

In this section we prove convergence of $y(t)$ in the DD model with eligibility trace for matching shoulders reward structures. All results in this subsection apply to the system (1)–(3), (7) and (14)–(16). As in the analysis for the WSLS model, we consider the general case (17) when $y^* \notin \mathcal{Y}$. Also like the analysis for the WSLS model, the results generalize to nonlinear curves; however, for clarity of presentation we specialize to intersecting linear reward curves defined by

$$\begin{aligned} f_A(y) &= k_A y + c_A \\ f_B(y) &= k_B y + c_B \end{aligned} \tag{28}$$

where $k_A < 0$, $k_B > 0$ and $c_A, c_B > 0$.

We first look at the case when the subject does not switch choice at a given time t_0.

Lemma 9. *For any* $t_0 > 0$, *if* $y(t_0 - 1) < 1$ *and* $x_1(t_0 - 1) = x_1(t_0) = A$, *then there exists a finite* $t_1 \geq t_0$ *such that* $x_1(t) = A$ *for all* $t_0 \leq t \leq t_1$ *and* $y(t_1) = y(t_0 - 1) + 1/N$.

Proof of Lemma 9. If $x_N(t_0 - 1) = B$, then $y(t_0) = y(t_0 - 1) + 1/N$ and so the conclusion holds for $t_1 = t_0$. If on the other hand $x_N(t_0 - 1) = A$, then $y(t_0) = y(t_0 - 1)$. From (16) we know that $\phi_A(t_0 - 1) = \phi_A(t_0) = 1$ and $\phi_B(t_0) = 0$. Then it follows from (14) that $\omega_A(t_0 + 1) = r(t_0) = f_A(y(t_0)) = f_A(y(t_0 - 1)) = \omega_A(t_0)$ and from (15) that $\omega_B(t_0 + 1) = \omega_B(t_0)$. Since $x_1(t_0 - 1) = x_1(t_0) = A$, from (7) it must be true that $\omega_A(t_0) > \omega_B(t_0)$. Thus we know $\omega_A(t_0+1) > \omega_B(t_0+1)$, so again from (7), we have $x_1(t_0+1) = A$. In fact, the choice of A will be repeatedly chosen as long as the value of x_N remains A. However, since $y(t_0 - 1) < 1$, there must exist $t_1 \leq t_0 + N$ such that $x_N(t_1 - 1) = B$ and then for the same t_1, we have $x_1(t) = A$ for all $t_0 \leq t \leq t_1$, and $y(t_1) = y(t_0 - 1) + 1/N$. □

Using a similar argument, one can prove the following lemma which is the counterpart to Lemma 9.

Lemma 10. *For any* $t_0 \geq 0$, *if* $y(t_0 - 1) > 0$ *and* $x_1(t_0 - 1) = x_1(t_0) = B$, *then there exists a finite* $t_1 \geq t_0$ *such that* $x_1(t) = B$ *for all* $t_0 \leq t \leq t_1$, *and* $y(t_1) = y(t_0 - 1) - 1/N$.

Lemmas 9 and 10 imply that if $Z \in \{A, B\}$ is repeatedly chosen, then the anticipated reward for choice Z decreases as a result of the change in y while the anticipated reward for the alternative $\bar{Z}$ stays the same because the eligibility trace $\phi_{\bar{Z}}$ remains zero. Hence, a switch of choices is likely to happen after a finite time. Now we look at the case when the subject switches choice at time $t_0 > 0$, namely $x_1(t_0) = \bar{x}_1(t_0 - 1)$. Then from (16), we have $\phi_{x_1(t_0-1)}(t_0) = \epsilon$; correspondingly from update rules (14) and (15), we have $\omega_{x_1(t_0-1)}(t_0+1) = \omega_{x_1(t_0-1)}(t_0)+\epsilon(r(t_0)-\omega_{\bar{x}_1(t_0-1)}(t_0)) = \omega_{x_1(t_0-1)}(t_0) + \epsilon(\omega_{\bar{x}_1(t_0-1)}(t_0 + 1) - \omega_{\bar{x}_1(t_0-1)}(t_0))$. Hence, the magnitude of ϵ is critical in updating the value of the anticipated reward when a switch of choices happens. It should be pointed out that in Section 3, to be consistent with the exponential decay rate for eligibility trace in [12], we have made an assumption that ϵ is a small number. This assumption can be stated by restricting the upper bound for the convex combination of points on the f_A and f_B lines.

Assumption 1 (*Restricted Convex Combination*). For $y \in \mathcal{Y}$,

$$(1 - \epsilon)\min\{f_A(y), f_B(y)\} + \epsilon \max\{f_A(y), f_B(y)\} < f_A(y^*) = f_B(y^*).$$

The following result states that under certain circumstances the subjects will not immediately follow a switch of choice with another switch of choice.

Lemma 11. *Suppose Assumption 1 is satisfied. For any $t_0 > 0$, if $x_1(t_0) = A$, $x_1(t_0 - 1) = B$ and $y(t_0 - 1) < y^*$, then there exists a finite $t_1 \geq t_0$ such that $x_1(t) = A$ for all $t_0 \leq t \leq t_1$, and $y(t_1) > y^*$.*

Proof of Lemma 11. From (16) we know that $\phi_A(t_0) = 1$ and $\phi_B(t_0) = \epsilon$. So from (14) and (15), it follows that

$$\omega_B(t_0) = r(t_0 - 1) = f_B(y(t_0 - 1)), \tag{29}$$

$$\omega_A(t_0 + 1) = r(t_0) = f_A(y(t_0)), \tag{30}$$

and

$$\begin{aligned}\omega_B(t_0 + 1) &= \omega_B(t_0) + \epsilon(r(t_0) - \omega_A(t_0)) \\ &= \omega_B(t_0) + \epsilon(f_A(y(t_0)) - \omega_A(t_0)).\end{aligned} \tag{31}$$

Since $x_1(t_0) = A$, from (7) it must be true that $\omega_A(t_0) \geq \omega_B(t_0)$. Combining with (31), we have

$$\begin{aligned}\omega_B(t_0 + 1) &\leq \omega_B(t_0) + \epsilon(f_A(y(t_0)) - \omega_B(t_0)) \\ &= (1 - \epsilon)\omega_B(t_0) + \epsilon f_A(y(t_0)).\end{aligned}$$

Substituting (29), we have

$$\omega_B(t_0 + 1) \leq (1 - \epsilon)f_B(y(t_0 - 1)) + \epsilon f_A(y(t_0)).$$

Since $x_1(t_0) = A$ and $x_1(t_0 - 1) = B$, we know $y(t_0) = y(t_0 - 1)$ or $y(t_0) = y(t_0 - 1) + 1/N$. Since $y(t_0 - 1) < y^*$, it must be true that either $y(t_0) = y^u$ or $y(t_0) \leq y^l$. We consider these two cases separately. Case (a): $y(t_0) = y^u$. Set $t_1 = t_0$, then $y(t_1) > y^*$ holds trivially. Case (b): $y(t_0) \leq y^l$. Then $f_A(y^*) < f_A(y(t_0)) \leq f_A(y(t_0 - 1))$. Also,

$$\omega_B(t_0 + 1) \leq (1 - \epsilon)f_B(y(t_0 - 1)) + \epsilon f_A(y(t_0 - 1)) < f_A(y^*),$$

where the last inequality follows from Assumption 1. Combining these with (30) we know that

$$x_1(t_0 + 1) = A \tag{32}$$

and consequently $\phi_A(t_0 + 1) = 1$ and $\phi_B(t_0 + 1) = 0$. In fact, A will be repeatedly chosen, ϕ_A and ϕ_B will remain one and zero respectively until some finite time $t_1 > t_0$ for which $\omega_A(t_1) = r(t_1) = f_A(y(t_1))$ is less than or equal to $\omega_B(t_0 + 1)$ or $y(t_1) = 1$. Since $\omega_B(t_0 + 1) < f_A(y^*)$, it follows that $y(t_1) > y^*$. So we have proved the conclusion for case (b) and the proof is complete. □

Using a similar argument, one can prove the following lemma which is the counterpart to Lemma 11.

Lemma 12. *Suppose Assumption 1 is satisfied. For any $t_0 > 0$, if $x_1(t_0) = B$, $x_1(t_0 - 1) = A$ and $y(t_0 - 1) > y^*$, then there exists a finite $t_1 \geq t_0$ such that $x_1(t) = B$ for all $t_0 \leq t \leq t_1$, and $y(t_1) < y^*$.*

As simulations and reported experiments indicate, the deterministic DD model fits subjects' behavior only when ϵ is bounded away from zero. Hence, we make the following assumption:

Assumption 2 (*Bounded ϵ*). The positive number ϵ is bounded below, satisfying

$$\epsilon \geq \max\left\{\frac{-k_A/N}{f_A(y^{l'}) - f_B(y^{l'})}, \frac{k_B/N}{f_B(y^{u'}) - f_A(y^{u'})}, \frac{f_A(y^{u'}) - f_B(y^{l'})}{f_A(y^l) - f_B(y^l)}, \frac{f_B(y^{l'}) - f_A(y^{u'})}{f_B(y^u) - f_A(y^u)}\right\}.$$

One consequence of Assumption 2 is that in certain scenarios, we can estimate the increment of the anticipated rewards.

Lemma 13. *Suppose Assumption 2 is satisfied. For any $t_0 > 0$, if $y(t_0) < y^{l'}$, $x_1(t_0 - 1) = x_1(t_0) = B$ and $x_1(t_0 + 1) = A$, then $\omega_B(t_0 + 2) \geq \omega_B(t_0 + 1) - k_A/N$.*

Proof of Lemma 13. Since $x_1(t_0-1) = x_1(t_0) = B$ and $x_1(t_0+1) = A$, it follows that

$$\omega_B(t_0 + 1) \leq \omega_A(t_0 + 1) = \omega_A(t_0) < \omega_B(t_0) \tag{33}$$

where

$$\omega_B(t_0 + 1) = f_B(y(t_0)) < f_B(y(t_0 - 1)) = \omega_B(t_0). \tag{34}$$

Then

$$\begin{aligned}\omega_B(t_0 + 2) &= \omega_B(t_0 + 1) + \epsilon\,(\omega_A(t_0 + 2) - \omega_A(t_0 + 1)) \\ &\geq \omega_B(t_0 + 1) + \epsilon\,(\omega_A(t_0 + 2) - \omega_B(t_0)) \\ &= \omega_B(t_0 + 1) + \epsilon\,(f_A(y(t_0 + 1)) - f_B(y(t_0 - 1)))\,.\end{aligned}$$

From $x_1(t_0 - 1) = x_1(t_0) = B$ and $x_1(t_0 + 1) = A$, we know that $y(t_0) = y(t_0-1)$ or $y(t_0) = y(t_0-1)-1/N$ and $y(t_0+1) = y(t_0)$ or $y(t_0+1) = y(t_0)+1/N$. Since $y(t_0) < y^{l'}$, it follows that $y(t_0+1) \leq y^{l'}$ and $y(t_0 - 1) \leq y^{l'}$. Because of the monotonicity of f_A and f_B, it follows that $\omega_B(t_0 + 2) \geq \omega_B(t_0 + 1) + \epsilon(f_A(y^{l'}) - f_B(y^{l'}))$. Using $\epsilon \geq \frac{-k_A/N}{f_A(y^{l'})-f_B(y^{l'})}$ in Assumption 2, we reach the conclusion. □

Following similar steps and using $\epsilon \geq \frac{k_B/N}{f_B(y^{u'})-f_A(y^{u'})}$ in Assumption 2, one can prove the following lemma which is the counterpart to Lemma 13.

Lemma 14. *Suppose Assumption 2 is satisfied. For any $t_0 > 0$, if $y(t_0) > y^{u'}$, $x_1(t_0 - 1) = x_1(t_0) = A$ and $x_1(t_0 + 1) = B$, then $\omega_A(t_0 + 2) \geq \omega_A(t_0 + 1) + k_B/N$.*

It is critical to examine the time instances at which the subject makes a switch from one choice to another. Thus let $\mathcal{T}$ denote the set of time instances for which $t > 0$ is in $\mathcal{T}$ if and only if $x_1(t) \neq x_1(t - 1)$. We are also interested in studying some subsets of $\mathcal{T}$. Define $\mathcal{T}_A \triangleq \{t : t \in \mathcal{T}, x_1(t) = A, y(t - 1) < y^*\}$ and $\mathcal{T}_B \triangleq \{t : t \in \mathcal{T}, x_1(t) = B, y(t - 1) > y^*\}$.

As in the analysis of the WSLS model, we consider $x_i(0)$, $i = 1, \ldots, N$, where $0 < y(0) < 1$. Then $\omega_{x_1(0)}(1) = r(0) = f_{x_1(0)}(y(0))$ is determined correspondingly. To simplify the analysis and rule out degenerate cases, we make the following assumption about the value of $\omega_{\bar{x}_1(0)}(1)$.

Assumption 3 (*Bounded* $\omega_{\bar{x}_1(0)}(1)$). Let $0 < y(0) < 1$ and $\omega_{x_1(0)}(1) = f_{x_1(0)}(y(0))$. The initial value $\omega_{\bar{x}_1(0)}(1)$ satisfies

$$\omega_{\bar{x}_1(0)}(1) < \omega_{x_1(0)}(1), \tag{35}$$

$$\omega_{\bar{x}_1(0)}(1) < f_A(y^*), \tag{36}$$

and

$$\omega_{\bar{x}_1(0)}(1) > \max\{f_A(1), f_B(0), f_A(1) + (k_B + k_A)/N, f_B(0) - (k_B + k_A)/N\}. \tag{37}$$

Now we are ready to study how the DD model with the eligibility trace evolves with time.

Lemma 15. *Suppose all the Assumptions 1–3 are satisfied. Then,*

$$\mathcal{T} = \mathcal{T}_A \cup \mathcal{T}_B.$$

Proof of Lemma 15. From (35) we know that $x_1(1) = x_1(0)$ and in fact $x_1(0)$ will be repeatedly chosen until the reward for choosing $x_1(0)$ is below $\omega_{\bar{x}_1(0)}(1)$. Such a switch will always happen because of (37). Let t_0 denote the time at which such a switch happens, i.e., $x_1(t_0 - 1) = x_1(0)$ and $x_1(t_0) = \bar{x}_1(0)$. In view of (36), we know that $y(t_0 - 1) < y^*$ if $x_1(0) = B$ and $y(t_0 - 1) > y^*$ if $x_1(0) = A$. So $t_0 \in \mathcal{T}_A$ if $x_1(0) = B$ and $t_0 \in \mathcal{T}_B$ if $x_1(0) = A$, and thus $t_0 \in \mathcal{T}_A \cup \mathcal{T}_B$. By inspection, if $y(t_0 - 1) \in \mathcal{L}$, then either Lemma 9 or Lemma 10 is applicable to t_0; if on the other hand, $y(t_0 - 1) \notin \mathcal{L}$, then either Lemma 11 or Lemma 12 is applicable to t_0. This implies that $x_1(t_0)$ will be repeatedly chosen such that the time of the next switch $t_1 \geq t_0$ satisfies $t_1 \in \mathcal{T}_{\bar{x}_1(t_0)} \subset \mathcal{T}_A \cup \mathcal{T}_B$. Further, Lemmas 9–12 can be applied again to t_1. Then, by induction, we know all time instances for which a switch happens belong to $\mathcal{T}_A \cup \mathcal{T}_B$. □

As becomes apparent later, when the DD model with the eligibility trace converges, the values of y^*, k_A and k_B affect the range of the interval containing y^* to which $y(t)$ converges. In this paper, we are interested in the sufficient conditions under which such an interval is $\mathcal{L}'$.

Assumption 4 is concerned with the relative positions of points on the reward lines f_A and f_B corresponding to values of y in $\mathcal{L}'$.

Assumption 4 (*Points in* $\mathcal{L}'$).

$$f_B(y^l) + \epsilon(f_A(y^u) - f_B(y^u)) \geq f_A(y^{u'}), \tag{38}$$

$$f_A(y^u) + \epsilon(f_B(y^l) - f_A(y^l)) \geq f_B(y^{l'}). \tag{39}$$

Proposition 2. *Suppose all the Assumptions 1–4 are satisfied. If* $t_0 \in \mathcal{T}$, $y(t_0 - 1) \in \mathcal{L}'$, *then* $y(t) \in \mathcal{L}'$ *for all* $t \geq t_0$.

Proof of Proposition 2. The conclusion can be proved by induction if we can prove the following fact: There is always a finite $t_1 > t_0$ such that $t_1 \in \mathcal{T}$ and $y(t) \in \mathcal{L}'$ for all $t_0 \leq t \leq t_1$. In order to prove this fact, we need to consider four cases.

- Case (a): $x_1(t_0) = A$ and $y(t_0 - 1) = y^{l'}$. Then

$$\begin{aligned}\omega_B(t_0 + 1) &> f_B(y^{l'}) + \epsilon\left(f_A(y(t_0)) - f_B(y^l)\right) \\ &\geq f_B(y^{l'}) + \epsilon\left(f_A(y^l) - f_B(y^l)\right),\end{aligned}$$

  where the first inequality holds since $f_B(y^l) > f_B(y^{l'})$ and the last inequality holds because $f_A(y(t_0)) \geq f_A(y^l)$. In view of the inequality $\epsilon \geq \frac{f_A(y^{u'}) - f_B(y^{l'})}{f_A(y^l) - f_B(y^l)}$ in Assumption 2 and combining with Lemma 11, we know that there is always a finite $t_1 > t_0$ such that $t_1 \in \mathcal{T}_B$ and $y(t) \in \mathcal{L}'$ for all $t_0 \leq t \leq t_1$.
- Case (b): $x_1(t_0) = A$ and $y(t_0 - 1) = y^l$. Then

$$\begin{aligned}\omega_B(t_0 + 1) &\geq f_B(y^l) + \epsilon\left(f_A(y(t_0)) - \omega_A(t_0)\right) \\ &> f_B(y^l) + \epsilon\left(f_A(y^u) - f_B(y^u)\right),\end{aligned}$$

  where the first inequality holds because $\omega_A(t_0) \geq \omega_B(t_0)$ and the last inequality holds because $f_A(y(t_0)) \geq f_A(y^u)$ and $\omega_A(t_0) = \omega_A(t_0 - 1) < \omega_B(t_0 - 1) = f_B(y^*) < f_B(y^u)$. Then in view of (38), we know $\omega_B(t_0 + 1) \geq f_A(y^{u'})$, and so there is always a finite $t_1 > t_0$ such that $t_1 \in \mathcal{T}_B$ and $y(t) \in \mathcal{L}'$ for all $t_0 \leq t \leq t_1$.
- Case (c): $x_1(t_0) = B$ and $y(t_0 - 1) = y^u$. Following similar steps as in Case (b) and using (39), we know there is always a finite $t_1 > t_0$ such that $t_1 \in \mathcal{T}_A$ and $y(t) \in \mathcal{L}'$ for all $t_0 \leq t \leq t_1$.
- Case (d): $x_1(t_0) = B$ and $y(t_0 - 1) = y^{u'}$. Following similar steps as in (a) and using the inequality $\epsilon \geq \frac{f_B(y^{l'}) - f_A(y^{u'})}{f_B(y^u) - f_A(y^u)}$ in Assumption 2 and combining with Lemma 12, we know that there is always a finite $t_1 > t_0$ such that $t_1 \in \mathcal{T}_A$ and $y(t) \in \mathcal{L}'$ for all $t_0 \leq t \leq t_1$.

In view of the discussion in Cases (a)–(d), we conclude that the proof is complete. □

Proposition 3. *Suppose all the Assumptions 1–4 are satisfied. There is always a finite* $T \in \mathcal{T}$ *for which* $y(T - 1) \in \mathcal{L}'$.

Proof of Proposition 3. From (37) in Assumption 3, we know that $x_1(0)$ cannot be repeatedly chosen more than N times, and thus there is $t_0 \in \mathcal{T}_A \cup \mathcal{T}_B$. If $y(t_0 - 1) \in \mathcal{L}'$, then set $T = t_0$ and we reach the conclusion. If on the other hand, $y(t_0 - 1) \notin \mathcal{L}'$, then either Lemma 13 or Lemma 14 applies to $t_0 - 1$. Without loss of generality, suppose Lemma 13 applies, then $\omega_B(t_0 + 1) \geq \omega_B(t_0) - k_A/N$. By (37) $\omega_B(t_0 + 1) > f_A(1)$. So, there exists a t_1 which is the smallest element in $\mathcal{T}_B$ such that $t_1 > t_0$. Then $\omega_A(t_1) \geq \omega_B(t_0 + 1) + k_A/N > \omega_B(t_0)$. In fact, one can check that switches will continue to exist and $\omega_{\bar{x}_1(t)}$ will be a monotonically strictly increasing function until $\omega_{\bar{x}_1(t)}$ reaches either $f_A(y^{u'})$ or $f_B(y^{l'})$ at some finite time T'. Set T to be the smallest element in $\mathcal{T}$ that is greater than T', then it must be true that $y(T - 1) \in \mathcal{L}'$. □

Combining Propositions 2 and 3, we have proved the main result of this section.

Theorem 3. *For system (1)–(3), (7), (14)–(17) and (28), if Assumptions 1–4 are satisfied, then there exists a finite* $T > 0$, *such that* $y(t) \in \mathcal{L}'$ *for all* $t \geq T$.

4.3. Discussion

We have proved in Theorems 2 and 3 for the WSLS decision-making model and the DD decision-making model that the proportion y of A's in the moving window of the N most recent decisions converges to the neighborhood $\mathcal{L}'$ of the matching point y^*. This means that eventually the decision maker confines decisions to those corresponding to the four values of y closest to y^*. However, convergence is more direct for the WSLS decision maker as compared to the DD decision maker. In the case of the WSLS decision maker, the very first time y reaches $\mathcal{L}$, it will remain in $\mathcal{L}'$ for all future time. This is because the WSLS makes its greedy decisions between A and B depending only on the two most recent awards. On the other hand, in the case of the DD decision maker, if the same choice is made repeatedly when y is in $\mathcal{L}'$ (and thus in $\mathcal{L}$), then y can pass through $\mathcal{L}'$ without getting trapped. It is only when the decision maker switches choice while $y \in \mathcal{L}'$ that y will remain in $\mathcal{L}'$ for all future time. The DD decision maker can persist with repeated choices even as rewards decrease because the expected reward for the alternative, which depends on the reward received the last time the alternative was chosen, can be relatively low.

In the psychology studies, N is chosen large enough to push the limits of what humans can remember. N can be interpreted as a measure of task difficulty since it determines how many of the past choices influence the present reward. N controls the resolution of the reward: reward depends on y, which takes values in the set $\mathcal{Y} = \{j/N, j = 0, 1, \ldots, N\}$. The convergence we have shown is tighter for a larger N, i.e., $\mathcal{L}$ and $\mathcal{L}'$ are smaller. Likewise, the convergence rate is slower for a larger N since it will take more choices to make the same magnitude change in y. The resolution also affects the sensitivity of decision making: peaks and dips in the reward curve of width smaller than $1/N$ won't be measured by the decision maker. For example, in the rising optimum reward (Fig. 1(b)), a small enough N would enable the decision maker to "jump over" the minimum in the f_A curve.

5. Application

As one possible application of the convergence results of Section 4, we formulate a human-supervised robotic foraging problem where the human makes sequential binary decisions. To make the theory applicable, we propose a map from the human-supervised robotic foraging problem to a two-alternative forced-choice task. We discuss conditions for such a map that would justify using results on human decision making to help guide the design of an integrated human–robot system. We are particularly interested in leveraging the matching behavior phenomenon since it is so strongly supported by psychology experiments and is formally proven in the previous section. However, we expect that the human supervisor will contribute to the complex foraging task in a variety of ways for which there may not be extensive insight nor formal models. Accordingly, we do not aim to use the existing models of human decision making to replace the human. The map proposed here is only a first step and we expect that alternatives will improve and extend our central idea of identifying and leveraging parallels between what human subjects do in psychology experiments and what human operators do in complex tasks. Our focus is on human-in-the-loop issues; examples of developed strategies useful for collective foraging include exploration [28], coverage [29] and gradient climbing [30].

The robotic foraging problem that we are interested in is partly motivated by the producer–scrounger (PS) foraging game that models the behavior of group-foraging animals [31,32]. This model has been successful in predicting animals' decisions either to look for food (produce) or to exploit the food found by other foragers (scrounge). Two results in the study of the PS game are especially relevant to the robotic foraging problem of interest. First, the rewards for producing and scrounging are functions of the proportion of scroungers in the animal group, and such reward curves are similar to the matching shoulders curves studied in Section 4 [31]. Second, scientists in a recent field study have introduced techniques to manipulate the reward curves in order to predictably shift the equilibrium of the group decision-making dynamics [32]. This motivating biological study suggests that for an integrated human–robot team, one may improve decision-making performance by adaptively changing how the reward is perceived by the team, e.g., by the robots or by the humans or by the robots and humans together.

Consider a team of N autonomous robots, foraging in a spatially distributed field $\mathcal{S}$, that are *remotely supervised* by a human. Each robot forages in one of two modes: when *exploring* the robot searches for regions of high density of resource and when *exploiting* it stays put in a high density region and collects resources. The role of the human supervisor is to make the choice for each individual robot, sequentially in time ($t = 0, 1, 2, \ldots$), as to whether it should explore or exploit. After each supervisor decision, the robot or the group of robots then provides the supervisor with a measure of performance. For example, after a decision to exploit, an estimate could be made of the amount of resource to be collected in the next time period under the assigned foraging mode allocation. This estimate could be made also after a decision to explore. Alternatively after a decision to explore, an estimate of the amount information to be collected in the next time period might be reported instead. In either case, the estimate represents the reward for the supervisor's decision at time t. By reading robots' estimations and making sequential decisions, the human supervisor allocates each of the N robots to foraging modes, one at a time, with the objective of maximizing total performance. The human continues to re-assign robots' foraging modes as long as necessary; for example, in a changing environment re-allocation may be critical. We note that in the case that the human operator gets information on different performance metrics with different units, e.g., if resource and information are reported for different operator choices, an important design question is how to scale one measure of performance relative to the other. The relative scaling will affect the decision making similarly to the way that modifying reward curves affects the decision-making dynamics of foraging animal groups [32].

The role of the human supervisor in this particular robotic foraging problem is to make a sequence of binary decisions, analogous to those of the psychology studies. The human-supervised robot foraging problem, however, is necessarily more complex than the task presented to the human subject. In particular, the rewards reported to the human supervisor will likely depend dynamically on the decision history. A preliminary numerical study of collectively foraging robots in fields of continuous, time-varying distribution of resource does show evidence of reward curves that, at times, have slopes akin to those near a matching point [33]. In the study, the reward includes measures of both estimated resource collected and estimated information collected. The measure of resource collected signals how well the exploiters are able to collect. The measure of information signals when explorers are doing a productive job searching out neglected regions with the expectation of finding new patches of dense resources. The study shows that growing numbers of exploiters are useful only up to a point at which their value drops off because there are not enough explorers to help direct them to high density patches. Likewise, growing numbers of explorers are useful only up to a point at which their value drops off because there are not enough exploiters to collect resource at the high density patches that have been discovered.

We show here a different, simpler numerical simulation of collective robot foraging to illustrate a reward curve for a particular decision sequence and compare it to the reward curves studied in [12,13]. Consider a planar $L \times L$ region $S = \{(u, v) | u \in [0, L], v \in [0, L]\}$ with distributed resources. Let the resource distribution with two big patches be described by the sum of Gaussians

$$\Phi(u, v) = a_1 \mathrm{e}^{-((u-\bar{u}_1)^2+(v-\bar{v}_1)^2)/\sigma_1^2} + a_2 \mathrm{e}^{-((u-\bar{u}_2)^2+(v-\bar{v}_2)^2)/\sigma_2^2} \quad (40)$$

where $(\bar{u}_i, \bar{v}_i)$ is the center of patch i, $i = 1, 2$, and a_i, σ_i are peak value and spread for patch i.

The explore and exploit control laws are heuristics and serve as proxies (for illustration) for more sophisticated foraging behaviors. Each exploiting robot collects resources at a high rate but does not move. Each exploring robot moves at a common constant speed and updates its heading $\theta_k(t)$ as

$$\begin{aligned}\theta_k(t+1) &= \theta_k(t) + \alpha_k(t)(\psi_k(t) - \theta_k(t)) \\ &\quad + (1 - \alpha_k(t))(\hat{\theta} + k_{\mathrm{com}}(\phi_k(t) - \theta_k(t))), \end{aligned} \quad (41)$$

$$\alpha_k(t) = \begin{cases} 1 & \text{if } \Phi(u_k(t), v_k(t)) > r^* \\ 0 & \text{otherwise.} \end{cases} \quad (42)$$

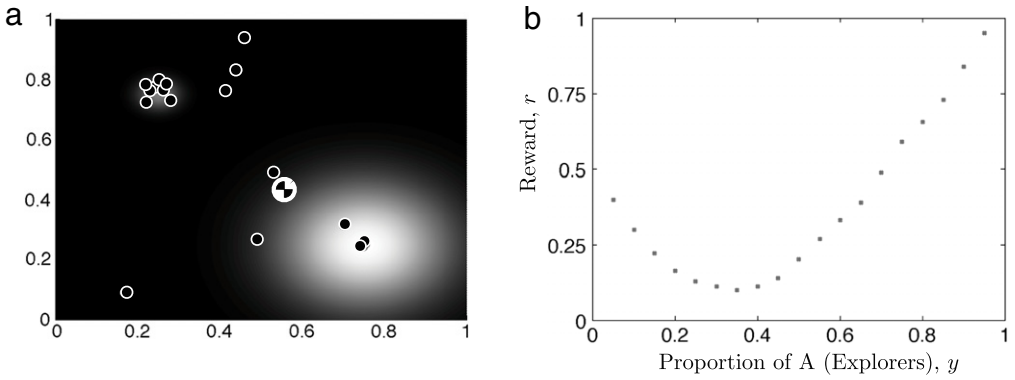


Fig. 4. Numerical foraging experiments with pre-defined allocation decision sequence. (a) Snapshot of resource field and robots (black circles). The large white circle is the center of mass of explorers. (b) Reward for choosing A (explorer) averaged over 100 simulations.

$\hat{\theta}$ is a random variable, ϕ_k is the direction of the vector from robot k to the center of mass of all exploring robots, ψ_k is the direction of the gradient of the resource at robot k's location and $k_{\text{com}} > 0$ is a constant. α_k switches value when robot k's measured resource exceeds threshold r^*. If $\alpha_k = 0$ exploring is a random walk plus an attraction to the center of mass of explorers; if $\alpha_k = 1$ exploring is gradient climbing on the resource. The resources collected by robot k at time t are

$$\rho_k(t) = \begin{cases} \gamma_A r(u_k(t), v_k(t)) & \text{if } k \text{ an explorer} \\ \gamma_B r(u_k(t), v_k(t)) & \text{if } k \text{ an exploiter} \end{cases} \tag{43}$$

where $\gamma_A, \gamma_B \in [0, 1]$ and $\gamma_A < \gamma_B$. A decision on the allocation of robots to exploring and exploiting is made every T units of time. After the decision is made at time t, the robots forage and then report the reward at time $t + T$ as the total resources collected during the interval $[t, t + T]$.

Fig. 4 shows the results of simulations with $N = 20$ robots for a pre-defined decision sequence. All 20 robots are initially exploiters located by a random uniform distribution in a disk of radius $0.1\,L$ about the center of patch 1. Every decision is to choose A, which is to assign the next robot to be an explorer. That is, $y(0) = 0$ and $y(kT) = 0.05k$, $k = 1, \ldots, 20$. The simulation was repeated 100 times. Fig. 4(a) shows a snapshot of the resource field and foraging robots in one run and Fig. 4(b) shows the reward f_A averaged over all 100 runs. The field parameters are $L = 1, a_1 = 0.5, \sigma_1 = 0.01, a_2 = 1.0$, and $\sigma_2 = 0.2$. Exploring robots have speed $v = 0.02$ and $T = 20$, $\gamma_A = 1/2$, $\gamma_B = 3/4$, $k_{\text{com}} = 0.009$ and $\hat{\theta}$ is taken from a uniform distribution in the interval $[-\frac{\pi}{6}, \frac{\pi}{6}]$.

The structure of f_A in Fig. 4 is similar to that of the rising optimum reward structure of Fig. 1(b). With few explorers, increasing the number of explorers means less resource since explorers abandon patch 1 and search regions that may have lower resource levels. However, with more explorers, increasing the number of explorers means greater success at finding the large resource peak at patch 2. The negative slope in the reward curve suggests the existence of a matching point, even possibly in the case of less structured decision sequences.

6. Concluding remarks

In this paper we consider two models for human decision making in the two-alternative forced-choice task. We provide a formal analysis of matching behavior, a human response strongly supported by psychology experiments that often corresponds to suboptimal decision strategies. We prove convergence to matching for a class of reward curves when the human decision maker is represented by the WSLS model and a deterministic limit of the DD model. We discuss the differences in the convergence of the WSLS decision maker as compared to the DD decision maker. As an application, we formulate a framework for a human-supervised robotic foraging problem, where the human supervisor makes decisions, based on a report of performance from the robots, that compare to the kinds of decisions made by the human subject in the psychology experiments, based on a computer generated reward. The psychology experiments can be viewed as an idealized representation of the more dynamic human-supervised foraging task.

In ongoing work, we are investigating probabilistic models that describe human decision-making dynamics. This framework will be useful when considering multiple human decision makers along with the results discussed in [15]. We are also exploring how to design an integrated human–robot system to our advantage. We aim to bring to bear the computational power of the robotic system without replacing the human supervisor. As an example, we have been studying adaptive laws for the robot feedback that use only local information but help the human make optimal decisions.

Acknowledgements

We have benefited greatly from discussions with Philip Holmes, Jonathan Cohen, Damon Tomlin, Andrea Nedic, Pat Simen and Deborah Prentice. We thank the anonymous reviewers for their insights and constructive comments. This research was supported in part by AFOSR grant FA9550-07-1-0-0528 and ONR grant N00014-04-1-0534.

References

[1] R. Murphey, P.M. Pardalos (Eds.), Cooperative Control and Optimization, Springer, 2002.
[2] V. Kumar, N. Leonard, A.S. Morse (Eds.), Cooperative Control, Springer, 2005.
[3] P. Antsaklis, J. Baillieul (Eds.), IEEE Transactions on Automatic Control: Special Issue on Networked Control Systems, vol. 49, IEEE, 2004, p. 9.
[4] P. Antsaklis, J. Baillieul (Eds.), Proceedings of the IEEE: Special Issue on Technology of Networked Control Systems, vol. 95, IEEE, 2007, p. 1.
[5] R. Simmons, S. Singh, F. Heger, L. Hiatt, S. Koterba, N. Melchior, B. Sellner, Human–robot teams for large-scale assembly, in: Proc. NASA Science Technology Conference, 2007.
[6] T. Kaupp, A. Makarenko, Measuring human–robot team effectiveness to determine an appropriate autonomy level, in: Proc. IEEE Int. Conf. on Robotics and Automation, 2008.
[7] A. Steinfeld, T. Fong, D. Kaber, M. Lewis, J. Scholtz, A. Schultz, M. Goodrich, Common metrics for human–robot interaction, in: Proc. Human–Robot Interaction Conference, 2006.
[8] R. Alami, A. Clodic, V. Montreuil, E.A. Sisbot, R. Chatila, Task planning for human–robot interaction, in: Proc. Joint Conf. on Smart Objects and Ambient Intelligence, ACM, 2005.
[9] J.G. Trafton, N.L. Cassimatis, M.D. Bugajska, D.P. Brock, F.E. Mintz, A.C. Schultz, Enabling effective human–robot interaction using perspective-taking in robots, IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans 35 (4) (2005) 460–470.
[10] R. Herrnstein, Rational choice theory: Necessary but not sufficient, American Psychologist 45 (1990) 356–367.

[11] P.R. Montague, G.S. Berns, Neural economics and the biological substrates of valuation, Neuron 36 (2002) 265–284.
[12] R. Bogacz, S.M. McClure, J. Li, J.D. Cohen, P.R. Montague, Short-term memory traces for action bias in human reinforcement learning, Brain Research 1153 (2007) 111–121.
[13] J. Li, S.M. McClure, B. King-Casas, P.R. Montague, Policy adjustment in a dynamic economic game, PLoS One e103 (2006) 1–11.
[14] D.M. Egelman, C. Person, P.R. Montague, A computational role for dopamine delivery in human decision-making, Journal of Cognitive Neuroscience 10 (1998) 623–630.
[15] A. Nedic, D. Tomlin, P. Holmes, D.A. Prentice, J.D. Cohen, A simple decision task in a social context: Preliminary experiments and a model, in: Proc. of the 47th IEEE Conference on Decision and Control, 2008, pp. 1115–1120.
[16] B. Donmez, M.L. Cummings, H.D. Graham, Auditory decision aiding in supervisory control of multiple unmanned aerial vehicles. Human Factors: The Journal of the Human Factors and Ergonomics (in press), OnlineFirst, published 2009 as doi:10.1177/0018720809347106.
[17] K.C. Campbell Jr., W.W. Cooper, D.P. Greenbaum, L.A. Wojcik, Modeling distributed human decision-making in traffic flow management operations, in: 3rd USA/Europe Air Traffic Management R&D Seminar, Napoli, 2000.
[18] R. Herrnstein, in: Howard Rachlin, David I. Laibson (Eds.), The Matching Law: Papers in Psychology and Economics, Harvard University Press, Cambridge, MA, USA, 1997.
[19] M. Cao, A. Stewart, N.E. Leonard, Integrating human and robot decision-making dynamics with feedback: Models and convergence analysis, in: Proc. of the 47th IEEE Conference on Decision and Control, 2008, pp. 1127–1132.
[20] L. Vu, K. Morgansen, Modeling and analysis of dynamic decision making in sequential two-choice tasks, in: Proc. of the 47th IEEE Conference on Decision and Control, 2008, pp. 1121–1126.
[21] H. Robbins, Some aspects of the sequential design of experiments, Bulletin of American Mathematical Society 58 (1952) 527–535.
[22] M. Nowak, K. Sigmund, A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game, Nature 364 (1993) 56–58.
[23] R. Bogacz, E. Brown, J. Moehlis, P. Holmes, J.D. Cohen, The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks, Psychological Review 113 (2006) 700–765.
[24] B.K. Oksendal, Stochastic Differential Equations: An Introduction with Applications, Springer-Verlag, Berlin, 2003.
[25] P. Simen, J.D. Cohen, Explicit melioration by a neural diffusion model, Brain Research 1299 (2009) 95–117.
[26] P.R. Montague, P. Dayan, T.J. Sejnowski, A framework for mesencephalic dopamine systems based on predictive Hebbian learning, Journal of Neuroscience 16 (1996) 1936–1947.
[27] R.S. Sutton, A.G. Barto, Reinforcement Learning, MIT Press, Cambridge, MA, 1998.
[28] D. Baronov, J. Baillieul, Reactive exploration through following isolines in a potential field, in: Proc. of 2007 American Control Conference, 2007, pp. 2141–2146.
[29] J. Cortes, S. Martinez, T. Karatas, F. Bullo, Coverage control for mobile sensing networks, IEEE Transactions on Robotics and Automation 20 (2004) 243–255.
[30] P. Ogren, E. Fiorelli, N.E. Leonard, Cooperative control of mobile sensor networks: Adaptive gradient climbing in a distributed environment, IEEE Transactions on Automatic Control 49 (2004) 1292–1302.
[31] L.A. Giraldeau, T. Caraco, Social Foraging Theory, Princeton University Press, Princeton, NJ, USA, 2000.
[32] J. Morand-Ferron, L.A. Giraldeau, L. Lefebvre, Wild carib grackles play a producer–scrounger game, Behavioral Ecology (2007) 916–921.
[33] C. Baldassano, N.E. Leonard, Explore vs. exploit: Task allocation for multi-robot foraging. Preprint. 2009. Available online http://www.princeton.edu/~naomi/publications.html.